

# 距離学習とデータ拡張による深層学習の汎化能力改善とそのディープフェイク検出応用

代表研究者

檜作 彰良

立命館大学理工学部 講師

## 1 はじめに

近年, Variational Auto-Encoder (VAE) や Generative Adversarial Network (GAN) などの深層学習 (Deep Learning) 技術の発展によって, 特定の人物の顔や動きを別の人物のものに置き換え, または誰かの声を模倣することが可能となっている. 深層学習技術によって生成される様々な偽造画像・動画は, “ディープフェイク” と呼ばれている. Web 上で確認されたディープフェイク動画は年々増加傾向であり, この対応が喫緊の課題となっている.

従来研究において, 深層学習技術を用いて動画を“Real (本物)”か“Fake (偽物)”に判定するネットワークが開発されている. しかし, 従来のディープフェイク検出法の問題点として, 未学習のデータセットに適用した場合, ネットワークの検知性能が低下する問題がある. そこで本研究では, ディープフェイク検出のネットワークの汎化能力 (未学習のデータに対して, ネットワークが正しい解を出力する能力) を向上させるために, 距離学習とデータ拡張技術を用いたディープフェイク検出法を提案する. 距離学習は, 同じクラス (Real/Fake) に属する動画像間の距離を近づけ, 異なるクラスの動画像間の距離を遠ざけるようにネットワークを学習させることで, 識別性の高いより本質的な特徴量を抽出する. また様々なディープフェイク画像に対応するために, SBI (Self-Blended Images) [1]を用いて, ネットワークの学習データを擬似的に増強する.

## 2 実験試料

実験試料として, Deepfake の公開データセットである FaceForensics++ (FF++) [2], Celeb-DF [3], Face Forensics In the Wild (FFIW) [4]を用いた. FF++には 1,000 本のオリジナル動画 (Real) とこれらの画像に 4 種類の操作手法 (Deepfakes, Face2Face, FaceSwap, NeuralTextures) を適用して生成された計 4,000 本のディープフェイク動画 (Fake) が含まれている. 本研究では, ネットワークの学習用に FF++データセット (オリジナル画像のみ), 評価用に Celeb-DF データセットから 518 動画 (Real : 178 動画, Fake : 340 動画), FFIW データセットから 500 動画 (Real : 250 動画, Fake : 250 動画) を用いている. 図 1 に, FF++と Celeb-DF に含まれている映像の例を示す.



Fig.1 FF++と Celeb-DF に含まれている映像例

### 3 方法

提案手法では、分類ネットワークとして EfficientNet-b4 を使い、同ネットワークの中間層から得られる複数解像度の特徴マップに対して、距離学習手法の一つである ArcFace を適用することで判別性能の向上を図る。さらに、画像合成に基づく学習データ生成手法である Self-Blended Images (SBI) を導入し、ネットワークの検知精度を高める。

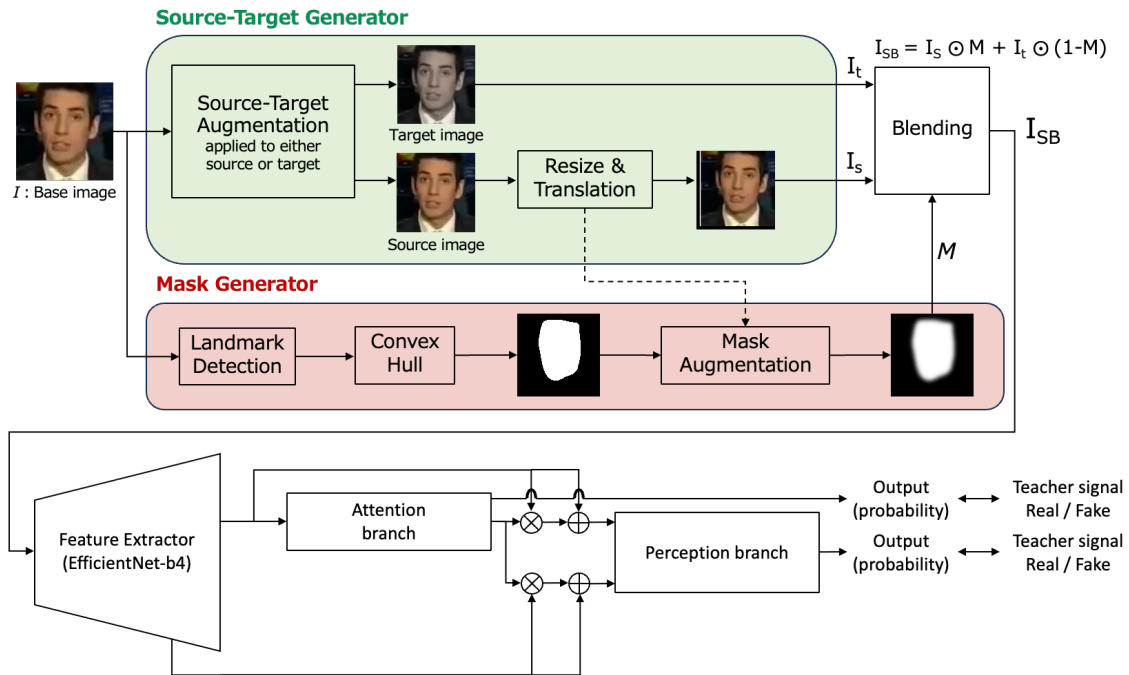


図1 提案ネットワークの処理の流れ

#### 3-1 SBI (Self-Blended Images)

図1に示すように、SBIはSource-Target GeneratorとMask Generatorより構成されている。Source-Target Generatorでは、入力された画像をSource-Target Augmentationにより、ソース画像とターゲット画像を生成する。そして、ソース画像に対して、色変換や並行移動などの変換を加える。一方、Mask Generatorでは、Dlib [5]を用いて取得した顔のランドマークに基づき、顔領域のマスク画像を生成する。最後に、マスク画像を用いて、ソース画像とターゲット画像をブレンドし、合成画像を生成する。

#### 3-2 EfficientNet-b4

EfficientNet-B4 [4]は、Tanらにより提案されたEfficientNetモデルの一つであり、ネットワークの深さ(depth)、幅(width)、および画像解像度(resolution)を同時にスケーリングするCompound Scalingによって設計された畳み込みニューラルネットワーク(CNN: Convolutional Neural Network)である。本ネットワークは、1つのStem、7つのBlock群、および最終層から構成されており、各Blockは複数のモジュールで構成されている。各モジュールは、Mobile Inverted Bottleneck Convolution (MBConv)を基本単位とし、Squeeze-and-Excitation (SE) モジュールを組み込むことで、特徴抽出を行っている。Stemでは、3×3の畳み込み層によって入力画像から特徴を抽出する。続く7つのBlockは、解像度やチャンネル数の異なるMBConvモジュールを段階的に適用する。各MBConvモジュールは、畳み込み演算、正規化、Swish活性化関数、およびチャネル間の関係性を強調するSEモジュールで構成される。最終層では、1×1の畳み込み層とグローバル平均プーリング、および全結合層を通じて、入力画像を”Real”、または”Fake”のいずれかに分類する。

### 3-3 Attention 機構

文献[7]でも用いられているように、本研究では、CNN が特定領域（本研究では顔領域）の特徴マップに注目して解析を行うことを可能とする Attention Mechanism を導入する．この Attention Mechanism では、顔領域に対して重み付けされた Attention Map を用いて、特徴マップ上の解析すべき領域を与える．提案ネットワークの Attention Branch は、2つの2次元畳み込み層（カーネルサイズ：3×3）、バッチ正規化層、ReLU 活性化関数、さらに1つの2次元畳み込み層（カーネルサイズ：1×1、チャンネル数：1）および Sigmoid 関数から構成される．EfficientNet-B4 の最終畳み込み層直前に出力される特徴マップ（カーネルサイズ：1×1）を Attention Branch に入力とし、最後の畳み込み層の出力に対して Sigmoid 関数を適用することで、Attention Map を生成する．ここで、Attention Map が顔領域に対応するように学習させるため、顔領域のマスク画像を教師信号として用いる．これらのマスク画像は、顔特徴点検出ライブラリ Dlib を用いて自動生成したものである．

### 3-4 ArcFace

Attention Map により重み付けされた特徴マップを全結合層へ入力する．ここで、全結合層の処理は、次式で表される．

$$\mathbf{Z}_i = \mathbf{W}_c \mathbf{x}_i + \mathbf{b}_c \quad (1)$$

但し、 $\mathbf{W}_c$  はクラス  $c$  における全結合層の重みベクトル、 $\mathbf{x}_i$  は入力データ  $i$  の特徴ベクトル、 $\mathbf{b}_c$  はバイアスである．全結合層において、重みベクトル  $\mathbf{W}_c$  と特徴ベクトル  $\mathbf{x}_i$  は正規化されている．ここで、バイアス  $\mathbf{b}_c$  を 0 としたとき、全結合層の処理は次のように変換することができる．

$$\mathbf{Z}_i = \|\mathbf{W}_c\| \|\mathbf{x}_i\| \cos \theta_c = \cos \theta_c \quad (2)$$

ここで、 $\theta_c$  は  $\mathbf{W}_c$  と  $\mathbf{x}_i$  のなす角である．学習時において、正解クラスの重みベクトル  $\mathbf{W}_c$  と特徴ベクトル  $\mathbf{x}_i$  とのなす角に angular margin penalty[8]である  $m$  を加える．これらを踏まえ、モデルの出力  $o(i)$  は、次式で定義される．

$$o(i) = \frac{\exp(s \cos(\theta_{y_i} + m))}{\exp(s \cos(\theta_{y_i} + m)) + \sum_{c \neq y_i} \exp(s \cos \theta_c)} \quad (3)$$

但し、 $y_i$  は入力データ  $i$  の正解クラス、 $s$  は Softmax 関数のスケールパラメータである．

### ネットワークの損失関数

提案ネットワークの損失関数は、次式で定義される．

$$L = -\frac{1}{N} \sum_{i=1}^N \log \frac{\exp(s \cos(\theta_{y_i} + m))}{\exp(s \cos(\theta_{y_i} + m)) + \sum_{t \neq y_i} \exp(s \cos \theta_t)} \quad (4)$$

ここで、 $N$  はデータ数である．Angular margin penalty ( $m$ ) を正解クラスの重みベクトルと特徴ベクトルとのなす角にのみ追加することで、それらの2つのベクトルがより近づくようにネットワークの学習が行われる．

### 評価指標

提案手法の評価指標としては、ROC 曲線 (Receiver Operating Characteristic curve) 下の面積である AUC (Area Under the Curve) を用いた．AUC の利点は、入力画像を Real と Fake に分類する際の閾値設定に依存せず、様々な閾値における識別性能を総合的に評価できる点にある．

## 4 結果

複数解像度の特徴マップに基づく分類の有用性を検証するために、最後の MBConv より出力された特徴マップのみを用いたネットワーク(network1)、最後及びその一つ前の MBConv より出力された特徴マップを用いたネットワーク(Proposed network)、最後及びその二つ前の MBConv より出力された特徴マップを用いたネットワーク(network2)を比較した．この結果、Proposed network の AUC(0.9442)が、その他のネットワーク(network1:0.9344, network2:

0.9374)よりも高い値が得られた。ここで、これらの結果は SBI により生成した画像で評価した。次に、ArcFace を用いたネットワーク(Proposed network with ArcFace)と用いないネットワーク(Proposed network without ArcFace)の精度比較を行った。その結果、Proposed network with ArcFace の AUC は 0.9459 であり、Proposed network without ArcFace (0.9442)よりも高い結果が得られた。通常の交差エントロピーを用いたネットワーク(Proposed network without ArcFace)の学習では、各クラス間の特徴空間におけるクラス内・クラス間の分布を考慮することは困難である。したがって、クラスが違うサンプルでも特徴空間で近接する可能性がある。一方、距離学習の一つである ArcFace を用いた提案ネットワーク(Proposed network with ArcFace)では、同じクラスに属するデータ間の距離を近づけ、異なるクラスのデータ間の距離を遠ざけるようにネットワークを学習させることで、識別性の高い特徴量を抽出できたと考える。

上記で決定した提案ネットワークを評価データセットに適用した結果、Celeb-DF における AUC は 0.9459、FFIW では 0.8259 を達成した。

## 5 まとめ

本研究では、分類モデルの一つである EfficientNet-b4 の中間層から得られる複数解像度の特徴マップに対して、距離学習手法の一つである ArcFace を適用することで判別性能の向上を行った。今後の課題として、提案ネットワークを様々なデータセットに適用し、提案ネットワークの有用性を評価する。

### 【参考文献】

- [1] K. Shiohara, T. Yamasaki, “Detecting deepfakes with self-blended images”, Proc. of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), 2022.
- [2] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, M. Nießner, “Faceforensics++: Learning to detect manipulated facial images”, Proc. of the IEEE/CVF international conference on computer vision (CVPR), pp. 1-11, 2019.
- [3] Y. Li, X. Yang, P. Sun, H. Qi, S. Lyu, “Celeb-DF: A large-scale challenging dataset for deepfake forensics”, Proc. of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp.3207-3216, 2022.
- [4] T. Zhou, W. Wang, Z. Liang, and J. Shen, “Face forensics in the wild”, Proc. of the IEEE/CVF international conference on computer vision (CVPR), pp.5778–5788, Nashville, TN, USA, June 2021.
- [5] M. Tan, Q. Le, Q., “Efficientnet: Rethinking model scaling for convolutional neural networks”, International conference on machine learning (PMLR), pp. 6105-6114, 2019.
- [6] D.E. King, “Dlib-ml: A machine learning toolkit”, The Journal of Machine Learning Research, 10, pp.1755-1758, 2009.
- [7] 綿引凌, 金子直史, 鷲見和彦, “Attention Map 修正による未学習のデータセットに対する Deepfake 検出性能向上”, 電子情報通信学会論文誌 A, Vol.106, No.12, pp.275-284, 2023.
- [8] J. Deng, J. Guo, N. Xue, S. Zafeiriou, S., “Arcface: Additive angular margin loss for deep face recognition”, Proc. of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp.4690-4699, 2019.

### 〈発表資料〉

| 題 名                     | 掲載誌・学会名等         | 発表年月        |
|-------------------------|------------------|-------------|
| 深層距離学習を用いた DeepFake 検出法 | フォーラム顔学 2024 抄録集 | 2024 年 11 月 |